DO CLOZE TESTS WORK?
OR, IS IT JUST AN ILLUSION?

JAMES DEAN BROWN
University of Hawai'i at Manoa

INTRODUCTION

Cloze procedure first appeared in the work of Wilson Taylor (1953), who studied the effectiveness of cloze as in instrument for assessing the relative readability of written materials for school children in the United States. Research then turned to the utility of cloze as a test reading proficiency among native speakers (for examples of this early work, see Bormuth 1965, 1967; Crawford 1970; Gallant 1965; or Ruddell, 1964). During the sixties, studies were also done on the value of cloze as a measure of overall ESL proficiency (for overviews of this early L2 cloze research, see Alderson 1978; Oller 1979; Cohen 1980).

As noted in Brown (1984), this literature on cloze as a measure of overall ESL proficiency has been far from consistent. The results of such studies, especially the reliability and validity results, have varied greatly both within and among studies. For example, the reliability estimates of the various cloze tests reported in the literature have ranged from .13 to .96 (Alderson 1979a; Bachman 1985; Brown 1980, 1983, 1984, 1988b, 1989, 1994; Brown, Yamashiro, & Ogane, 1999; Darnell 1970; Hinofotis 1980; Jonz 1976; Mullen, 1979; Oller 1972b; Pike 1973). Similarly, criterion-related validity coefficients have varied from .06 to .91 (Alderson, 1979a, 1980; Bachman, 1985; Brown, 1980, 1984, 1988b; Conrad, 1970; Darnell, 1970; Hinofotis, 1980; Irvine, Atai, & Oller, 1974; Mullen, 1979; Oller, 1972a & b; Oller & Inal, 1971; Revard, 1990; and Stubbs & Tucker, 1974).

Many of the studies cited in the previous paragraph were designed to explore ways to develop and interpret cloze tests in order to maximize their reliability and validity by manipulating following variables: (a) scoring methods, (b) frequency of deletions (e.g., every 5th, 7th, 9th word, etc.), (c) length of blanks, (d) text difficulty, (e) native versus

Second Language Studies, 21(1), Fall 2002, pp. 79-125.

non-native performance, and (f) number of items. In the process, some researchers made claims that cloze test items are primarily tapping students' abilities to manipulate linguistic elements at the clause or sentence level, as opposed to predominately focusing on intersentential elements (see, for instance, Alderson, 1979a; Markham, 1985; Porter, 1983). Other researchers claimed that cloze items measure predominantly at the intersentential level (see, for example, Bachman, 1985, Brown, 1983; Chavez-Oller, Chihara, Weaver, & Oller, 1985; Chihara, Oller, Weaver, & Chavez-Oller, 1977; and Jonz, 1987). All in all, the cloze research to date has been rather inconclusive with regard to reliability, validity, and even with regard to what cloze tests are measuring.

Background

This particular study is the next step in one strand of my research that stretches back over 20 years. During those two decades, I have been intrigued by cloze procedure because it often turns out to be a reasonably good test of overall English language proficiency, yet we really do not understand how it works. I will briefly review that strand of my own research in order to show how those studies led me to the research hypotheses that drive the present paper.

I began my quest to understand cloze testing with Brown (1980), which was clearly the work of an earnest, fledgling, and ignorant young researcher. In that paper, I examined the exact-answer, acceptable-answer, clozentropy, and multiple-choice scoring methods for scoring cloze tests and concluded that the exact-answer scoring method was probably the best overall. To my credit, I had the good sense to tell the readers that they should decide for themselves which to choose on the basis of whatever testing characteristics were the most important to them (from among usability, item discrimination, item facility, reliability, standard error of measurement, and validity).

That early paper was fundamentally flawed in my current view because, in interpreting the differences among four scoring methods, I failed to consider the distributions of scores and their effects on the relative values of my descriptive, item analysis, reliability, and validity statistics. For example, the acceptable-answer scoring results formed a near perfect distribution with the mean very well-centered (M = 25.58 out of 50) and room for almost exactly two standard deviations above and below that

mean (*SD* = 12.45). The other scoring distributions were all less well-centered with means above or below that mean and standard deviations that indicated non-normal distributions.

I have realized since publishing that study that my results might have turned out entirely differently had I chanced upon a passage that was either more difficult or easier than the one I did use, because the relative normality or skewedness of the four distributions would have been entirely different. Naturally, those differences in distributions would have affected the relative magnitudes of the item facility and discrimination values for the scores resulting from the four different scoring methods as well as the four sets of reliability and validity coefficients. More importantly, those differences would have altered my conclusions. Learning from the flaws of that study, I have since been keenly aware of the supreme importance of descriptive statistics and the distributions they represent in interpreting any statistical results. That realization will bear heavily on the results of the study reported in this paper.

In conducting the two studies reported in Brown (1983, and reprinted in Brown, 1994), I discovered among other things that the K-R21 estimate consistently provides a serious underestimate of the reliability of cloze tests (when compared to Cronbach alpha, K-R20, and other estimates of reliability) as shown in Table 1. I continued to calculate K-R21 in all of my subsequent cloze studies and compare the resulting estimates with alpha and K-R20 (for example, see Brown, 1993)—always getting aberrant results for K-R21, most of which were underestimates.

Table 1
K-R21 and Other Estimates of Cloze Test Reliability (adapted from Brown, 1983)

	EX Scoring		AC Scoring	
Reliability Estimate	GP 1	GP 2	GP 1	GP 2
Cronbach alpha	0.66	0.61	0.67	0.67
K-R20	0.64	0.60	0.67	0.67
Split-half adjusted	0.67	0.63	0.61	0.67
Flanagan's coefficient	0.66	0.63	0.61	0.67
Rulon's coefficient	0.66	0.63	0.61	0.67
K-R21	0.48	0.36	0.56	0.55

Continuing over the years to ponder the importance of these observations, it eventually dawned on me that I should turn to the original Kuder and Richardson (1937) article. In that article, I found that one fundamental difference between Kuder-Richardson formulas 20 and 21 is the assumption underlying K-R21 that items must be equal in difficulty. I realized that, while K-R21 could reasonably be expected to provide good estimates of reliability for typical multiple-choice tests where we revise by selecting items to be of similar difficulty (i.e., with IF values ranging form .30 to .70), such an expectation might not be equally tenable for a cloze test where numerous items are often very difficult (IF = .00) or sometimes even very easy (IF = 1.00). In my experience, the K-R21 coefficients for *multiple-choice* and other discrete-point tests were usually equal to or slight underestimates of Cronbach alpha or K-R20, while the K-R21 for cloze tests often produced serious underestimates for cloze tests. The hypothesis I have formed here is that these *serious underestimates of K-R21 might be accounted for by the fact that many cloze items violate the equal difficulty assumption.*

In Brown (1983), I also thought I learned that cloze blanks tend to provide a fairly representative sample of the language in the passages regardless of the starting point for the deletion pattern. It is, after all, reasonable to assume that even a semi-random sampling of words from a passage will be reasonably representative of the words in that passage (especially if there are sufficient blanks, as in a 50 item cloze test). However, at the same time I noticed, quite reasonably, that some items were testing at the sentential level while others were testing at the inter-sentential level. What I have come to realize since then is that only some of the items on a cloze test may be functioning well for a given population of students, so regardless of the fact that the blanks may provide a representative sample of the language in the passage, the variance produced by those items may only be coming from those few items that are functioning well. Thus, the test variance may not be representative of the sampled items, and in turn may not be representative of the passage. For that reason, I hypothesize here (as did Alderson, 1979b) that *samples of items that delete different words, even in the same passage, may produce cloze tests that are quite different.*

In Brown (1984), I returned to the issue of score distributions and the importance of

relative amounts of variance to the reliability and validity of cloze. Table 2 from that study shows cloze tests arranged from most widely dispersed scores (as indicated by the standard deviations and ranges) to least. The associated reliability values and validity coefficients appear to be directly related to the degree of dispersion as we would expect.

Table 2
Ranges of Talent in Relationship to Cloze Test Reliability and Validity (adapted from Brown, 1984)

Sample	*SD*	Range	Reliability Estimate	Validity Coefficient
1978a	12.45	46	0.95	0.90
1978b	8.56	33	0.90	0.88
1981a	6.71	29	0.83	0.79
D1981b	5.59	22	0.73	0.74
1982a	4.84	22	0.68	0.59
1982b	4.48	20	0.66	0.51
1982c	4.07	21	0.53	0.40
1982d	3.38	14	0.31	0.43

The important thing to note here is that all of the results in Table 2 are based on exactly the same cloze test administered to groups of students with varying ranges of ability in English. That study revealed that a given cloze test could simultaneously be one of the best cloze tests ever reported in the literature (i.e., reliability = .95 and validity = .90) and one of the worst (i.e., reliability = .31 and validity = .43) depending how well it fit the particular group of students involved. Based on both Brown (1983) discussed in the previous paragraph and Brown (1984) discussed in this paragraph, I hypothesize that *a sample of items that fits a group of high proficiency students may be quite different from the sample of items that fits a group of intermediate proficiency students; in other words, the items that are working well for students at different levels of proficiency may be quite different.*

The 1984 study also lead me to attempt in Brown (1988b) to systematically tailor the distribution of scores on a cloze test much like we typically do in revising multiple-choice or other types of discrete-point tests by selecting items that discriminate well. That

process worked reasonably well, increasing both the reliability and validity of the cloze tests even though the tailored version was developed from exactly the same passage as the original versions. I therefore hypothesize here that *a cloze test tailored for students at different proficiency levels may draw on different item types to achieve reasonable distributions and reliability.*

In Brown (1989), I began to wonder, in the course of developing and studying the 50 different cloze tests used later in Brown (1993), if it may not be true that many items are not functioning at all, usually because they are so difficult for students that every student is answering incorrectly or leaving them blank (but potentially because they are so easy that every student is answering correctly, a state of affairs that is probably more likely with native speakers).

In Brown (1989), I also argued that taking either the sentential or inter-sentential point of view to the exclusion of the other is absurd saying that:

> The point is that most linguists would concede that the English language is complex and is made up of a variety of constraints ranging at least from morphemic and clausal level grammar rules to discourse and pragmatic level rules of cohesion and coherence all of which interact in intricate ways. Based on sampling theory, it is also a safe assumption that semi-random selection procedures like those used in creating a cloze test will create a representative sample of whatever is being selected as long as the samples are large enough. This assumption is the basis of much of the research done in the world today. (p. 48)

I also pointed out in Brown (1989) that:

> The question appears to hinge on the degree to which words, i.e., the units being sampled in a cloze test, are constrained by all of the levels of rules that operate in the language. If there are indeed different levels operating in the language which constrain the choices of words that writers make and if semi-random sampling creates a representative selection of these words, there is no alternative but to conclude that cloze items tap a complex combination of morpheme to discourse level rules in approximately the same proportions as they exist in the language from which they were sampled. Thus taking either of the positions above (i.e., that cloze items are essentially sentential, or primarily intersentential) and then conducting studies to

> support either position is to insure that the investigators will find what they are looking for. If both types of constraints are in operation, then both schools of thought are correct in finding what they are looking for and fundamentally wrong in excluding the other possibility. (p. 48)

Having conducted the present study, I now look back on those words as partially correct and partially incorrect: I still did not understand that sampling an item and having that item actually contribute to the test score variance may be two entirely different things. In other words, I am here hypothesizing that *many cloze items may not be functioning at all in test variance terms even though they may be present in the test.*

In fact, in Brown (1993, results also discussed in Brown, 1998), I began to see and understand the effects of such non-functioning items, or "turned off" items, on the distributions of scores on the 50 cloze tests as well as on the reliability and validity statistics associated with those tests.

Purpose

The results presented in the present study are new analyses of the data used in two previous studies (Brown, Yamashiro, & Ogane, 1999, 2001). The 1999 study examined what happens when a cloze procedure is tailored for a group of Japanese students at a relatively high proficiency level, and the 2001 study did the same for a group of Japanese students with relatively low proficiency. It is in the context of these eight studies (the two in this paragraph and the six reviewed in the ***Background*** section) that I interpret the results presented in this paper.

In brief, the overall purpose of this project was to combine and reanalyze the data from Brown, Yamashiro, & Ogane (1999 and 2001) in order to explore what it is that makes items function well in a cloze test for students at different proficiency levels. To that end, the following research questions (based on the five italicized hypotheses in the previous section but in a different order) were investigated:

1. Are there significant differences between the five means and five variances produced by the different samples of items in the five cloze tests (a) when the five are administered to low and high proficiency students? (b) When they are scored using the EX and AC methods?

2. (a) How many of the 30 items in each of the ten cloze tests in this study are not functioning at all (i.e., have item facility values of zero)? (b) How many are outside the acceptable range of .30 to .70 for classical theory item facility? (c) How many are functioning poorly in item discrimination terms? (d) Do the results differ for different proficiency students?
3. To what degree does a relationship exist between the number of items falling outside the .30 to .70 range and the underestimation of reliability provided by K-R21? And, why should we care?

Because of the exploratory nature of this research, the alpha level for all statistical decisions was set at $\alpha < .05$.

METHOD

Participants

Two different groups of students were used in this study. One will be referred to as the high proficiency students and the other will be called the low proficiency students. These terms are only relative in the sense that one group is higher in proficiency on average than the other. Each will be described in turn.

High proficiency students. The high proficiency students were sampled from a very high-ranking private secondary school in Japan. Three first-year and three second-year "returnee" (students who have lived and studied overseas for at least two years) classes were selected to be in this group. Approximately 10% of the students were high-proficiency "regular" students who either volunteered for these returnee classes or were recommended by their teachers to do so. To put these classes in perspective, at the end of the school year, the Pre-TOEFL is administered to all first-year students. About two-thirds of the "returnee" students get perfect scores of 500 and the rest get scores in the middle-to-high 400s. Among the second-year students, who all take a regular institutional TOEFL, the returnee class students typically have an average of about 570, with scores ranging from the high 400s to the mid 600s. Of these students, 143 students took the cloze tests in this study.

Low proficiency students. The 193 low proficiency students were all in their second year in the Law and Political Science Department at Heisei International University. Most of the students had attended public schools in neighboring prefectures. Based on an informal biographical survey, no participants had lived overseas, and only about 10% had even visited an English-speaking country (for two weeks or less) prior to this study. About half the participants acknowledged that they did not like studying English. Most students want careers in local civil service jobs as police officers, government workers, etc. Hence, they might one day need some functional ability in English. No TOEFL scores or estimates were available for this group, but they were definitely of lower proficiency on the whole relative to the high proficiency students. That fact will be substantiated by the cloze test results that follow.

Materials

The initial five cloze passages used in this study were developed from a passage entitled "The Science of Automatic Control," which first appeared in Bachman (1985). Initially, five cloze tests were used in this study: Form A was the same as Bachman's original fixed-ratio cloze; Form B deleted the words one word to the right of the original deletion; Form C deleted the words two to the right of the original deletion; Form D deleted the words one to the left; and Form E deleted the words two to the left of the original deletion (example directions and the first ten items of Form E are shown in Appendix A; the exact-answer key for those first ten items is provided in Appendix B).

The cloze passages in this study were scored using both *exact-answer scoring* (only counts as correct that word which was originally deleted from the blank) and *acceptable-answer scoring* (scoring based on a glossary of possible answers for each blank as determined by native speakers). In this case, the glossary for acceptable-answer scoring was created by three teachers, who were native speakers of English; they generated a glossary for each item on each of the five forms before the tests were scored. Additional acceptable answers were necessarily added (based on agreement between two of the teachers) during the scoring phase of the project (as explained below); all previously scored tests were then checked again with the newly added acceptable answers included

in the glossary.

Procedures

The data used in this study are from the preliminary cloze test administrations at the two universities involved. These data represent the first round test development and administration of the pilot forms discussed in Brown, Yamashiro, & Ogane (1999, 2001) both of which provide additional information. The test development and administration of the pilot cloze tests included the following eight steps:

1. Five distinct cloze tests were developed on the basis of a single passage from Bachman (1985) (as described above in the ***Materials*** section).
2. The five forms were then photocopied and piled such that the five different cloze tests alternated in repeated patterns of A, B, C, D, and E.
3. During the test administration, the teachers distributed the cloze tests to students beginning at the top of the pile and proceeding to the bottom for each group of participants.
4. The participants were told in advance that their scores would not count in their course grades; they were then allowed fifteen minutes to complete the cloze tests.
5. Acceptable-answer keys (with a glossary of possible items for each item) were worked out by three native speakers of English before the tests were scored (as described above in the ***Materials*** section).
6. The pilot cloze tests were then scored for exact and acceptable answers.
7. The exact and acceptable answers were reported separately to the students.
8. All of the data were entered into a computer spreadsheet.

RESULTS

Descriptive Statistics

Table 3a shows descriptive statistics for the five cloze versions and two scoring methods for the low proficiency students ($N = 193$). Notice that, as would be expected, the means in Table 3a are generally very low for 30 item (k) tests, ranging for the exact-answer scoring from 0.72 to 3.50 and for the acceptable-answer scoring from 1.64 to 4.05. Notice also that, in half the cases, the standard deviations are larger than the means

and that in the other cases the standard deviations are almost as large as the means, all of which probably indicates positively skewed distributions. This interpretation is supported by the skew statistic which is positive in all cases and larger than two standard errors of skew (*ses*). The heights of the distributions also appear to be a problem for the low proficiency students, in some cases a major problem, as indicated by the high, in some cases very high, kurtosis values, all of which are positive and greater than two standard errors of kurtosis (*sek*). None of this is surprising given the relatively small ranges of scores (for a 30 point test) that themselves range from a low of 8 (0 to 7) to a high of 20 (0 to 19).

Table 3b shows the descriptive statistics for five cloze versions and two scoring methods for the high proficiency students ($N = 143$). Notice that, again as would be expected, the means in Table 3b are much higher than those in Table 3a, ranging for the exact-answer scoring from 7.79 to 11.44 and for the acceptable-answer scoring from 10.20 to 15.26. Notice also that all the standard deviations in Table 3b are higher than those in Table 3a, but also that, in all cases, there is enough room below and above the mean for two or three standard deviations, which apparently indicates normal distributions in all cases. This interpretation is supported by the skew statistic which is close to zero and less than two standard errors of skew (*ses*) in magnitude in all cases except *EX E*. The heights of the distributions also appear to be normal, as indicated by the relatively low kurtosis values, all of which are close to zero and less than two standard errors of kurtosis (*sek*) in magnitude in all cases except *EX E*. None of this is surprising given the relatively wide ranges of scores (for a 30 point test) that themselves range from a low of 14 (2 to 15) to a high of 20 (8 to 27).

Table 3a
Descriptive Statistics for Five Cloze Versions and Two Scoring Methods for the Low Proficiency Students (N = 193)

SCORING *FORM*	*M*	*SD*	Low	High	Range	Skew	*ses*	Kurtosis	*sek*	*k*	*n*
EXACT											
EX A	1.58	2.26	0	13	14	3.52	.39	16.47	.77	30	40
EX B	2.53	2.27	0	10	11	1.25	.40	1.89	.79	30	38
EX C	0.72	1.32	0	7	8	3.18	.39	13.19	.78	30	39
EX D	1.53	1.67	0	7	8	1.51	.40	2.57	.79	30	38
EX E	3.50	3.22	0	19	20	3.18	.40	14.27	.79	30	38
ACCEPTABLE											
AC A	1.80	2.33	0	13	14	3.08	.39	13.17	.77	30	40
AC B	2.92	2.45	0	10	11	1.08	.40	0.78	.79	30	38
AC C	1.64	1.71	0	7	8	1.03	.39	0.98	.78	30	39
AC D	2.13	1.76	0	7	8	1.18	.40	1.33	.79	30	38
AC E	4.05	3.32	0	19	20	2.52	.40	10.38	.79	30	38

Table 3b
Descriptive Statistics for Five Cloze Versions and Two Scoring Methods for the High Proficiency Students (N = 143)

SCORING FORM	*M*	*SD*	Low	High	Range	Skew	*ses*	Kurtosis	*sek*	*k*	*n*
EXACT											
EX A	8.59	3.41	1	16	16	-0.12	.45	0.30	.91	30	29
EX B	9.33	3.76	3	20	18	0.38	.45	0.20	.89	30	30
EX C	9.59	3.89	4	18	15	0.22	.45	-0.97	.91	30	29
EX D	7.79	3.11	2	15	14	-0.01	.46	-0.14	.93	30	28
EX E	11.44	3.39	6	22	17	0.99	.47	2.54	.94	30	27
ACCEPTABLE											
AC A	10.90	4.12	2	19	18	-0.35	.45	-0.08	.91	30	29
AC B	12.23	3.86	5	22	18	-0.27	.45	0.20	.89	30	30
AC C	14.76	5.19	6	24	19	-0.05	.45	-1.26	.91	30	29
AC D	11.32	4.35	4	22	19	0.65	.46	1.06	.93	30	28
AC E	15.26	4.22	8	27	20	0.89	.47	1.45	.94	30	27

ANOVR Analysis of Overall Results

An overall repeated-measures analysis of variance (ANOVR) procedure was run with cloze scores as the dependent variable and three independent variables: proficiency levels (Low & High), scoring methods (EX & AC), and forms (A, B, C, D, & E) to determine what if any overall main effects and interactions were significant.

Assumptions of ANOVR. Before running the ANOVR, the assumptions were checked. The skew statistics especially in Table 3a indicate that some of the distributions for some of the forms were skewed. This would appear to be a clear violation of the

assumption of normality that underlies ANOVR. However, these violations probably do not pose a serious problem. As Tabachnick and Fidell (2001) put it (citing Mardia, 1971): "A sample size that produces 20 degrees of freedom for error in the univariate case should ensure robustness of the test, as long as the sample sizes are equal… Even with unequal *n* and only a few DVs, a sample size of 20 in the smallest cell should ensure robustness [to violations of the assumption of normality]."

The data were also checked for univariate outliers (one was found, the same case, in the EX and AC scores) and multivariate outliers (five were found to be over $\chi^2 = 13.816$; $df = 2$; $p < .001$ using the Mahalonabis distance statistic provided in the Regression module of SPSS). The two scores for the one univariate outlier were lowered to match the next highest score within the distribution and the multivariate outliers were eliminated from the data. The ANOVR procedures in this study were run using these new data, as well as the original data. Since the results of both runs were very similar (i.e., no statistical decisions turned out to be different, and *p* values, eta squared, and power estimates were very similar), these univariate and multivariate outliers were viewed as having only minimal effect. Hence, the analyses of the original data are presented here because they are easier to interpret and understand.

The distributions were also checked for equal variances. Levene's test indicated that there were significant differences ($p < .01$) in error variances somewhere in this proficiency levels by scoring methods by forms design. Box's *M* statistic was also significant at $p < .01$. The F_{max} statistic also turned out to be significant ($F_{max} = 15.459$, $p < .01$) for the overall design. In an effort to isolate where the significant differences in variances were located, the F_{max} statistic was calculated across forms within proficiency levels and scoring methods. F_{max} turned out to be significant for the low proficiency students when their tests were scored by the EX or AC methods ($F_{max} = 5.96$, $p < .01$ and $F_{max} = 3.77$, $p < .01$, respectively), but not for the high proficiency students for either scoring method. Within the EX scoring method for low students, the F_{max} statistic turned out to be significant for four out of the ten possible pairings of variances at $p < .01$ and for eight out of the ten possible pairings at $p < .05$. Within the EX scoring method for low students, the F_{max} statistic turned out to be significant for four out of the ten possible pairings of variances at $p < .01$ and for eight out of the ten possible pairings at $p < .05$.

Within the EX scoring method for low students, the F_{max} statistic across forms turned out to be significant for four out of the ten possible pairings of variances at $p < .01$ and for eight out of the ten possible pairings at $p < .05$. Within the AC scoring method for low students, the F_{max} statistic across forms turned out to be significant for two out of the ten possible pairings of variances at $p < .01$ and for eight out of the ten possible pairings at $p < .05$. Clearly then, this design contained unequal variances. According to Tabachnick and Fidell (2001, p. 395), "If sample sizes are equal, evaluation of homogeneity of variance-covariance matrices is not necessary." The cell sizes in this design were approximately equal ranging from 27 to 40. Nonetheless, following the advice of Tabachnick and Fidell (2001), the impact of unequal cell sizes in this repeated measures design was assessed by running both SPSS Type I and III designs; again, since the results of both runs were very similar (i.e., no statistical decisions turned out different, and *p* values, eta squared, and power estimates were very similar), the default Type III results are reported here.

The ANOVR analysis. The ANOVR source table shown in Table 4 indicates that the main effects for scoring method, proficiency, and form were all significant at $p < .01$, as were all of their interaction effects. Thus there were non-chance mean differences between the EX and AC scoring methods, the low and high proficiency students, and the five forms of the cloze test.

Table 4

ANOVR Source Table for Scores by Proficiency Levels (Low & High), Scoring Methods (EX & AC), and Forms (A, B, C, D, & E)

Source	*SS*	*df*	*MS*	*F*	*p*	Partial Eta sq	Power
Within-Participants Effects							
Scoring	685.32	1	685.32	924.38	0.000	0.739	1.00
Scoring x Proficiency	370.90	1	370.90	500.28	0.000	0.605	1.00
Scoring x Form	59.69	4	14.92	20.13	0.000	0.198	1.00
Scoring x Proficiency x Form	22.65	4	5.66	7.64	0.000	0.086	1.00
Error (Within-Participants)	241.69	326	0.74				
Between-Participants Effects							
Proficiency	12946.15	1	12946.15	689.44	0.000	0.679	1.00
Form	700.30	4	175.07	9.32	0.000	0.103	1.00
Proficiency x Form	252.33	4	63.08	3.36	0.010	0.040	0.85
Error (Between-Participants)	6121.58	326	18.78				

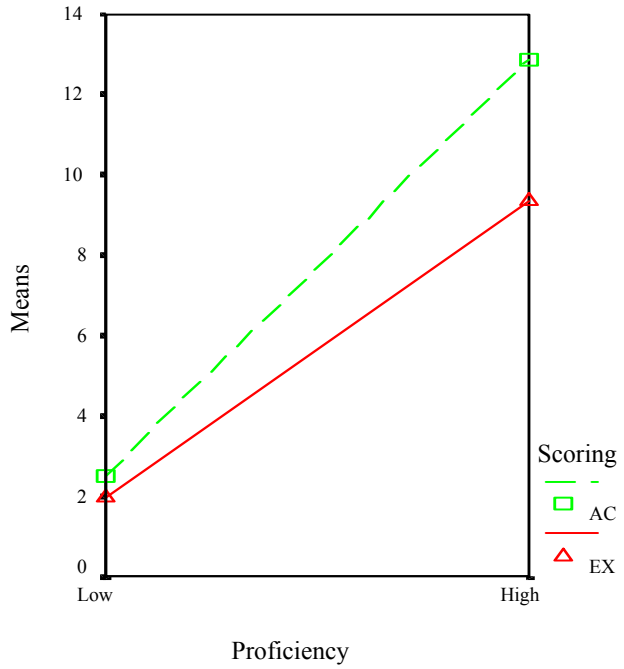

Figure 1: The Scoring by Proficiency Interaction

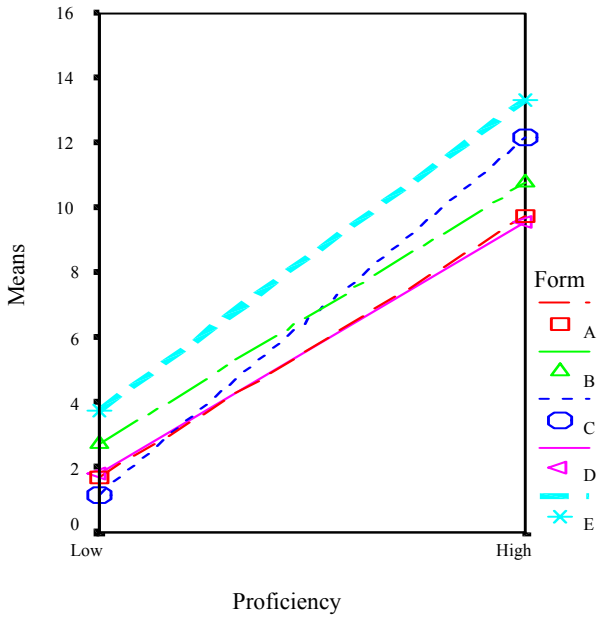

Figure 2: The Form by Proficiency Interaction

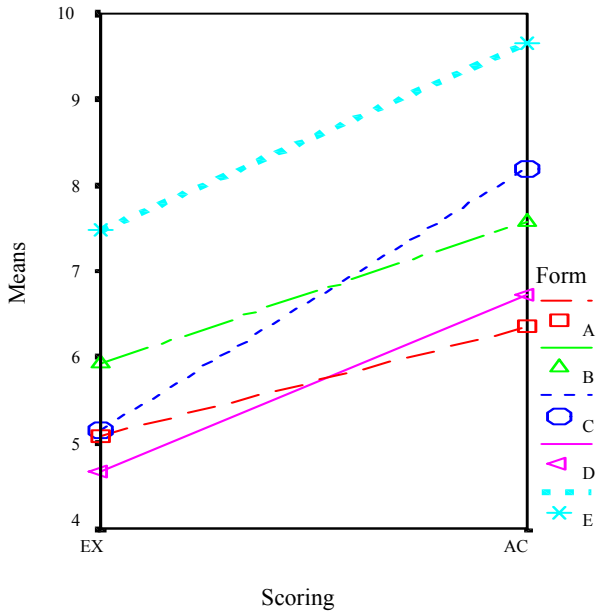

Figure 3: The Scoring by Form Interaction

In addition, the differences between means for proficiency, scoring, and forms were non-systematic as shown in Table 4 by the significant effects for all possible interactions and illustrated in Figures 1, 2, and 3. Figure 1 shows the significant interaction effect for

scoring method by proficiency level, and indicates that, quite naturally, both proficiency groups scored higher when the AC scoring was applied, but also that, on average, the high proficiency students benefited more from the use of the AC scoring than did the low proficiency students. Figure 2 shows the significant interaction effect for cloze form by proficiency level, and indicates that the five significantly different forms were not systematically different with regard to proficiency level. Most notably form C had the lowest mean for the low proficiency students, but was second from the highest for the high proficiency students, all of which resulted in the dotted line crossing three of the other lines in Figure 2. Similarly, form D crossed form A, but less sharply. Thus it appears that forms C and D give somewhat more advantage to the high proficiency students than the other forms do. Figure 3 shows the significant interaction effect for form by scoring method, and indicates that the five significantly different forms were not systematically different with regard to scoring method. Notice that again forms C and D end up crossing the other forms indicating that they result in higher scores for AC scoring relative to EX scoring than the other forms do.

Follow-up one-way ANOVAs across forms (within each of the scoring methods in each of the proficiency groups) and Scheffé tests indicated (all at $p < .01$) that: (a) forms A, C, and D were each significantly different from E for the low proficiency students using the EX scoring method; (b) forms A and C were significantly different from E for the low proficiency students using the AC scoring method; (c) D was significantly different from E for the high proficiency students using the EX scoring method; and (d) form A was significantly different from E for the high proficiency students when using the using the AC scoring method.

Reliability Statistics

Table 5a shows the reliability statistics for five cloze versions and two scoring methods for the low proficiency students, and Table 5b shows the same statistics for the high proficiency students. Both tables give the number of items (k), Cronbach alpha reliability (α), the Kuder-Richardson formula 21 (K-R21) reliability, and the standard error of measurement (SEM) based on Cronbach α reliability estimate. Notice in both tables that the K-R21 estimates are consistently lower than the Cronbach α estimates.

Table 5a
Reliability Statistics for Five Cloze Versions and Two Scoring Methods for the Low Proficiency Students

SCORING ***FORMS***	*k*	α	K-R21	SEM (α)
EXACT				
EX A	30	0.762	0.731	1.10
EX B	30	0.664	0.569	1.32
EX C	30	0.641	0.617	0.79
EX D	30	0.559	0.496	1.11
EX E	30	0.793	0.726	1.46
ACCEPTABLE				
AC A	30	0.738	0.712	1.19
AC B	30	0.657	0.580	1.43
AC C	30	0.525	0.486	1.18
AC D	30	0.478	0.374	1.27
AC E	30	0.769	0.706	1.60

Table 5b
Reliability Statistics for Five Cloze Versions and Two Scoring Methods for the High Proficiency Students

SCORING ***FORMS***	*k*	α	K-R21	SEM (α)
EXACT				
EX A	30	0.645	0.489	2.03
EX B	30	0.757	0.564	1.86
EX C	30	0.751	0.588	1.94
EX D	30	0.640	0.418	1.87
EX E	30	0.636	0.397	2.04
ACCEPTABLE				
AC A	30	0.738	0.612	2.11
AC B	30	0.715	0.532	2.06
AC C	30	0.832	0.747	2.13
AC D	30	0.768	0.649	2.09
AC E	30	0.718	0.599	2.24

The third column of numbers in Table 5c shows that, for the low proficiency students, the differences in reliability range from .031 to .095, or 3.1% to 9.5%, when using EX scoring, and from .026 to .104, or 2.6% to 10.4%, when using the AC scoring. The same column in Table 5d shows that, for the high proficiency students, the differences are much greater, ranging from 15.6% to 23.9% for the EX scoring, and from 8.5% to 18.3% when using the AC scoring. These results are similar to the effects I have found in previous studies, where I noticed serious and) consistent underestimates when using K-R21 (more about this topic in the **DISCUSSION** section).

Table 5c

Reliability and Item Facility for Five Cloze Versions and Two Scoring Methods for the Low Proficiency Students

SCORING *FORM*	α K-R21	Reliability Difference	Mean IF	Low IF	High IF	IF Range	IF .00	IF .03-.29	IF .30-.70	IF .71-1.00
EXACT										
EX A	.762	.731	.031	.05	.00	.28	.28	12	18	0	0
EX B	.664	.569	.095	.08	.00	.53	.53	13	14	3	0
EX C	.641	.617	.024	.02	.00	.18	.18	17	13	0	0
EX D	.559	.496	.063	.05	.00	.32	.32	18	11	1	0
EX E	.793	.726	.067	.12	.00	.53	.53	7	19	4	0
All 150 items				.034	.00	.53	.53	67	75	8	0
ACCEPTABLE										
AC A	.738	.712	.026	.06	.00	.28	.28	9	21	0	0
AC B	.657	.580	.077	.10	.00	.53	.53	11	16	3	0
AC C	.525	.486	.039	.05	.00	.28	.28	10	20	0	0
AC D	.478	.374	.104	.07	.00	.47	.47	16	12	2	0
AC E	.769	.706	.063	.14	.00	.53	.53	5	20	5	0
All 150 items				.084	.00	.53	.53	51	89	10	0

Table 5d

Reliability and Item Facility for Five Cloze Versions and Two Scoring Methods for the High Proficiency Students

SCORING ***FORMS***	α K-R21	Reliability Difference	Mean IF	Low IF	High IF	IF Range	IF .00	IF .03-.29	IF .30-.70	IF .71-1.00
EXACT										
EX A	.645	.489	.156	.29	.00	.76	.76	7	11	12	0
EX B	.757	.564	.193	.31	.00	.93	.93	9	7	11	3
EX C	.751	.588	.163	.32	.00	.93	.93	8	6	13	3
EX D	.640	.418	.222	.26	.00	.86	.86	11	8	7	4
EX E	.636	.397	.239	.38	.00	.93	.93	6	6	13	5
All 150 items				.312	.00	.93	.93	41	38	56	15
ACCEPTABLE										
AC A	.738	.612	.126	.36	.00	.86	.86	7	9	8	6
AC B	.715	.532	.183	.41	.00	.97	.97	6	7	12	5
AC C	.832	.747	.085	.49	.00	.93	.93	3	7	12	8
AC D	.768	.649	.119	.38	.00	.93	.93	5	8	13	4
AC E	.718	.599	.119	.51	.04	.93	.89	3	4	15	8
All 150 items				.430	.00	.93	.93	24	35	60	31

Item Analysis

One of the most interesting results of this study is that, generally speaking, the item statistics indicate that the cloze tests involved were not very good norm-referenced tests. If these tests had been typical multiple-choice tests, they would need much revision before I would be willing to use them for making decisions. Yet the results in the previous section indicate that they are at least moderately reliable. How can we reconcile these facts? Let's begin by examining the mean item facility values obtained for each cloze test when using EX and AC scoring, then examine the mean item discrimination estimates.

Item facility. Table 5c also shows item facility statistics for the five cloze versions and two scoring methods for the low proficiency students. Notice that the mean IF values shown in the fourth column of numbers are very low, as low as .02 for *EX C*, but even the

highest, .14 for *AC E*, is very low, indicating that these cloze tests were very difficult for the students involved with only 14% answering correctly on average across items on the easiest test. However, also note that the items varied considerably in difficulty: not a single student was able to answer some items as shown by the Low IF values of .00 (reported in the fifth column of numbers in Table 5a) and up to 53% of the students were able to answer other items as indicated by the High IF values of .53 for *EX B*, *EX E*, *AC B*, and *AC E*. The last three columns of numbers in that table show the frequency of items in three item difficulty ranges. Note that between 25 and 30 of the items on these cloze tests had IF values between .00 and .29, indicating that virtually all of the items were difficult. An additional 0 to 5 items fell in the more moderate difficulty range of .30 to .70, and no items fell in the range between .71 and 1.00. Typically, when developing norm-referenced tests, test designers hope to keep items ranging in IF from .30 to .70. Thus, from an IF perspective alone, most of these cloze items are not effective for these low proficiency students regardless of scoring method.

Table 5d shows IF values for the same five cloze test and two scoring methods, but this time for the high proficiency students. Naturally, the mean IF values shown in the fourth column of numbers are not as low as they were for the low proficiency students: ranging instead from a low of .26 for *EX D* to .51 for *AC E*. These values are still fairly low, indicating that these cloze tests were somewhat difficult for the students involved with only 51% answering correctly on average across items on even the easiest of the tests. However, again the items varied considerably in difficulty: not a single student was able to answer some items as shown by the Low IF values of .00 (reported in the fifth column for all tests except *AC E*) and up to 97% of the students were able to answer other items as indicated by the High IF values of .97 for *AC B*. The last three columns of numbers in that table show the frequency of items in three item difficulty ranges. Note that 7 to 19 of the items on these cloze tests had IF values between .00 and .29. An additional 7 to 15 items fell in the more moderate difficulty range of .30 to .70, and 0 to 8 items fell in the range between .71 and 1.00. Clearly, many more of the items are falling in the range of .30 to .70 that test designers would like to have on a norm-referenced test. Thus, from an IF perspective alone, these cloze items are much more appropriate for these high proficiency students than they were for the low proficiency students.

However, a substantial number of items still fall outside the IF range that I would expect from effective items on a norm-referenced multiple-choice test.

Item discrimination. The first column of numbers in Table 6a shows the average item discrimination values for each of the five cloze tests and two scoring methods for the low proficiency students. Notice that they range from .07 to .17 for the EX scoring and somewhat higher from .12 to .19 for the AC scoring. The next six columns show frequencies (the first three columns) and percentages (the last three columns) of items that had the best ID values of .30 or more, had weak ID estimates between .01 and .29, or were completely switched off with ID values of .00 (in this case, meaning that nobody answered them correctly, although the same .00 could result from everybody answering an item correctly). For both the EX and AC scoring methods combined, only between 2 (6.67%) and 10 (33.33%) of the items were discriminating well at above .30, while 7 (23.33%) to 16 (53.33%) were discriminating in a weak manner, and 5 (16.67%) to 18 (60.00%) were contributing nothing at all to the test variance because nobody was answering them correctly. I call these last items *switched off* because they were doing absolutely nothing, that is, they contributed nothing to the means or item variances. Thus, most of the items on these cloze tests (66.67% to 93.34%) were either weak discriminators or completely switched off when administered to the low proficiency students. Such a high proportion of items that do not discriminate well would never be tolerated on a well-developed norm-referenced multiple-choice test.

Table 6a

Item Discrimination for Five Cloze Versions and Two Scoring Methods for the Low Proficiency Students

SCORING FORM	Mean ID	Frequency *Best* Items (ID = .30+)	Frequency *Weak* Items (ID .01-.29)	Frequency *Turned Off* (ID = .00)	Percentage *Best* Items (ID = .30+)	Percentage *Weak* Items (ID .01-.29)	Percentage *Turned Off* (ID = .00)
EXACT							
EX A	.12	5	13	12	16.67	43.33	40.00
EX B	.15	7	10	13	23.33	33.33	43.33
EX C	.07	2	11	17	6.67	36.67	56.67
EX D	.11	5	7	18	16.67	23.33	60.00
EX E	.17	8	15	7	26.67	50.00	23.33
All 150 items	.124	27	56	67	18.00	37.33	46.67
ACCEPTABLE							
AC A	.13	5	16	9	16.67	53.33	30.00
AC B	.16	9	10	11	30.00	33.33	36.67
AC C	.12	6	14	10	20.00	46.67	33.33
AC D	.12	5	9	16	16.67	30.00	53.33
AC E	.19	10	15	5	33.33	50.00	16.67
All 150 items	.144	35	64	51	23.33	42.67	34.00

Table 6b

Item Discrimination for Five Cloze Versions and Two Scoring Methods for the High Proficiency Students

SCORING FORM	Mean ID	Frequency *Best* Items (ID = .30+)	Frequency *Weak* Items (ID .01-.29)	Frequency *Turned Off* (ID = .00)	Percentage *Best* Items (ID = .30+)	Percentage *Weak* Items (ID .01-.29)	Percentage *Turned Off* (ID = .00)
EXACT							
EX A	.23	12	11	7	40.00	36.67	23.33
EX B	.26	16	5	9	53.33	16.67	30.00
EX C	.29	14	8	8	46.67	26.67	26.67
EX D	.23	12	7	11	40.00	23.33	36.67
EX E	.22	11	13	6	36.67	43.33	20.00
All 150 items	.246	65	44	41	43.33	29.33	27.33
ACCEPTABLE							
AC A	.28	17	6	7	56.67	20.00	23.33
AC B	.26	15	9	6	50.00	30.00	20.00
AC C	.38	20	7	3	66.67	23.33	10.00
AC D	.30	14	11	5	46.67	36.67	16.67
AC E	.29	14	13	3	46.67	43.33	10.00
All 150 items	.302	80	46	24	53.33	30.67	16.00

Similarly, the first column of numbers in Table 6b shows the average item discrimination values for the high proficiency students. Notice that they are generally higher ranging from .22 to .29 for the EX scoring and even higher from .26 to .30 for the AC scoring. For both scoring methods combined, more items were discriminating well for the high proficiency students with between 11 (36.67%) and 20 (66.67%) of the items discriminating at above .30, while 5 (16.67%) to 13 (43.33%) were discriminating in a weak manner, and 3 (10.00%) to 11 (36.67%) were contributing nothing at all to the test variance because nobody was answering them correctly. Thus, even though these cloze tests were working substantially better for the high proficiency students than they did for the low proficiency students, a large number of the items (between 33.33% to 63.33%) were not discriminating very well or not at all when administered to the high proficiency students. Such a high proportion of items that do not discriminate well would typically not be tolerated on a norm-referenced multiple-choice test.

Items functioning well for low and high proficiency students. Given that the item analysis showed relatively poor item discrimination results across the board, the next question that arises is which items are discriminating. More to the point, the above results led me to wonder if it was the same or different items that were functioning well in the two groups. So I looked at the item level results overall in terms of how many items were discriminating and how many of those items were the same or unique for the low and high proficiency students.

Among other things, Appendixes C and D show that those items that were discriminating for the low and high groups were not exactly the same. In fact, the proportion of the discriminating items that was *unique* (i.e., discriminating with one group but not the other) ranged from 24% to 69% depending on the form involved, the scoring method, and of course, the group. In short, the two groups were receiving substantially different tests because different items within the pool of all cloze items were functioning well for the two groups.

DISCUSSION

In this section, I will directly address the research questions that were posed at the outset of this study. Those research questions will serve as subheadings to help guide the reader.

1. Are there significant differences between the five means and five variances produced by the different samples of items in the five cloze tests (a) when the five are administered to low and high proficiency students? (b) When they are scored using the EX and AC methods?

In direct answer to the first research question, Table 4 showed significant mean differences in scores for the five forms of the cloze test when they were administered to low and high proficiency students and when they were scored using EX and AC methods. In addition, all possible interactions were significant, indicating that the observed significant main effects were not 100% systematic as shown in Figures 1, 2, and 3. Furthermore, one-way ANOVAs and follow-up Scheffé tests run separately across the five forms for the two proficiency groups and two scoring methods showed exactly where those differences lay (all detailed differences significant at $p < .01$ were between some other form and form E).

The F_{max} statistic also showed numerous statistically significant differences ($p < .01$) between pairs of variances across the 20 cells of the proficiency levels (Low & High) by scoring methods (EX & AC) by forms (A, B, C, D, & E) ANOVR design. In addition, follow-up analyses showed statistically significant differences ($p < .01$) for numerous pairs of variances among the five forms in each proficiency group and each scoring method.

Given the classical definition of parallel forms (equal means, equal variances, and equal covariances), the cloze tests in this study indicate cases of non-parallel forms for at least two of those criteria in each and every set of five forms. In other words, selecting new starting points and creating five forms of a cloze test then administering those cloze tests to randomly selected groups of students does not appear to support the notion that those five forms are indeed parallel and equivalent, regardless of the scoring method used

(EX or AC) or proficiency level of the groups (low or high).

2. (a) How many of the 30 items in each of the ten cloze tests in this study are not functioning at all (i.e., have item facility values of zero)? (b) How many are outside the acceptable range of .30 to .70 for classical theory item facility? (c) How many are functioning poorly in item discrimination terms? (d) Do the results differ for different proficiency groups?

As shown in Tables 5c and d, many items are not functioning well at all. Depending on the proficiency group, scoring method, and form, between 3 (10%) and 18 (60%) of the items in these 30 item cloze tests are not functioning at all, i.e., they are contributing doing nothing at all to either the item variance or test variance as indicated by IF values of .00.

As also shown in Tables 5c and d, most of the items are outside the acceptable range of .30 to .70 for classical theory item facility for the low proficiency students. More precisely, 142 out of 150 items (or 94.6%) were unacceptably easy or difficult when using the EX scoring, and 140 out of 150 items (or 93.3%) were unacceptably easy or difficult when using the AC scoring. Similarly, for the high proficiency students, 94 out of 150 items (or 62.7%) were unacceptably easy or difficult when using the EX scoring, and 90 out of 150 items (or 60.0%) were unacceptably easy or difficult when using the AC scoring.

From a simple IF point of view, then, these cloze tests do not appear to be functioning very well as classical theory norm-referenced tests. It is hard to imagine any tester finding those statistics acceptable for any operational multiple-choice test. Perhaps, given the moderately high reliabilities in most cases, the item discrimination estimates are relatively high despite the fact that most of the items are too easy or too difficult for the students.

Table 6a shows that (c) the most of the items are functioning poorly in terms of item discrimination for the low proficiency (82% of the EX scored items are turned off or weak and 76.67% of the AC items are the same). Table 6b shows that many of the items are also functioning poorly in terms of item discrimination for the high proficiency (56.67% of the Ex scored items are turned off or weak and 46.67% of the AC items are

the same).

From both the IF and ID points of view, then, these cloze tests do not appear to be functioning very well as classical theory norm-referenced tests. Again, it is hard to imagine any tester finding those statistics acceptable for any operational multiple-choice test. Why are we willing to accept them for cloze tests? It may be that we have been blinded by the fact that such cloze tests appear to be reasonably reliable.

3. To what degree does a relationship exist between the number of items falling outside the .30 to .70 range and the underestimation of reliability provided by K-R21? And, why should we care?

Early in this paper, I hypothesized that "serious underestimates of K-R21 might be accounted for by the fact that many cloze items violate the equal difficulty assumption." For Tables 5c and 5d combined, the Pearson correlation coefficient between the reliability differences shown in the third column of numbers and the ranges shown in the seventh column is .83, which indicates a substantial amount of overlap about 69% ($.83^2 = .6899 \approx .69$) between the amount of variation in item facility values and the differences found between Cronbach α and K-R21. This relationship may be due to the lesser and greater violations of the equal item difficulty assumption that underlies the K-R21 statistic (Kuder & Richardson, 1937).

CONCLUSIONS

To sum up, then, let's reconsider the original hypotheses that I raised (in italics) throughout the ***Background*** section near the top of this paper. It turns out that, based on the results of this study, they all appear to be true:

1. Serious underestimates of K-R21 do seem to be accounted for by the fact that many cloze items violate the equal difficulty assumption.
2. Samples of items that delete different words, even in the same passage, do appear to produce cloze tests that are quite different.
3. A sample of items that fits a group of high proficiency students does seem to be quite different from the sample of items that fits a group of intermediate

proficiency students; in other words, the items that are working well for students at different levels of proficiency appear to be quite different

4. A cloze test tailored for students at different proficiency levels does seem to draw on substantially different item types to achieve reasonable distributions and reliability.
5. Many cloze items do not appear to be functioning at all in test variance terms even though they may be present in the test.

All along, then, the K-R21 underestimates have been trying to tell me that there is something wrong with the way cloze tests function. I think I can now characterize cloze tests as a test development technique, wherein we more or less randomly develop items in sufficient numbers so that, even though many of them do not function well, the test appears to be at least moderately reliable.

Since the items are developed from contextualized language, we have been willing to accept this situation in favor of what we assumed to be a valid sample of the items in the universe of all possible items in the (written) English language. Unfortunately, depending on the proficiency level of the students and the range of talent involved, as well as on the scoring method employed, many, in some cases, most of those items may not be functioning at all because they are completely turned of, or are at a best functioning poorly with IF levels outside the .30-.70 range and low discrimination indices.

As a corollary, the fact that many items are doing nothing will tend to mean that the passage is too difficult (at least when scored EX answer) for the students involved. The fact that many items are doing nothing might also explain why K-R21 is often a serious underestimate of the reliability of a cloze test. This fact might explain, in turn, why K-R20, Cronbach alpha, etc. are almost always high for cloze tests: if many items are switched off they create a false pattern of consistency across items, false in the sense that it is consistency that isn't discriminating or creating any sort of useful score information. Such items create the sense that students are all the same (i.e., none of them know these switched off items).

So just what is it that we now know about cloze items, how they work, and what makes them different from other tests? On the one hand, cloze tests do *not* appear to work well at all:

1. Unless a passage of the correct level of difficulty is found, the test will be made up largely of items that are switched off.
2. Because many items are likely to be switched off or poor discriminators, we should begin the construction of a cloze test with a very large number of items.
3. Even so, many items contribute nothing to test variance.
4. Even so, the students may be experiencing great frustration because many items are difficult or impossible for them to answer.
5. The items that are functioning for one high proficiency group may not function at all for another low proficiency group, or even for another high proficiency group.
6. Cloze tests administered to students of different ability levels are automatically testing different things because only those items that at least some of the students can answer will discriminate.
7. Because we do not know what cloze is doing, the items that are switched off have no meaning (that is, we do not learn what students cannot do in the same sense we might on a multiple-choice diagnostic grammar test or vocabulary test) because we do not know why they answered incorrectly.

On the other hand, cloze tests are marvelously adaptive.

1. Cloze tests are based on contextualized written language.
2. It is not difficult to get people to try taking a cloze test because of the human need to fill gaps (closure) which appears to be almost a compulsion among students.
3. As I pointed out in Brown (1986), students must predict in a manner similar to how they must predict in the reading process (if Goodman, 1967, and Smith, 1978, are even partially right).
4. A cloze passage that is in the ballpark difficulty-wise with enough items may serve to spread students out who are in very different ranges of ability.
5. In a sense, cloze challenges students with a semi-random selection of language items.
6. Students will then only correctly answer those items at their level of ability.
7. Hence, cloze tests administered to students of different ability levels will automatically be testing different things because only those items that at least some of the students can answer will discriminate.

So, should we continue using cloze tests? Given that the pros and cons are fairly even in number in the two lists above, I would say it is not yet necessary to throw the baby out with the bath water. There is much that is good about cloze tests and much that we can still learn from using them. Sure, we must be cautious in how we interpret the scores on cloze tests, and naturally, additional research is always necessary. Nonetheless, we now know how cloze tests adapt themselves to the ability levels of the particular group of students being tested: those items that do not discriminate are by-and-large switched off leaving mostly items that do discriminate to contribute to the test variance.

However, I am not sure we should continue letting the cards fall where they will by selecting every n^{th} word in developing cloze tests. The results of this study would seem to indicate that the every n^{th} word strategy is far too inefficient for responsible use in decision-making. Instead, we should probably use what we now know about the way some cloze items discriminate (and most others do not) to refine the strategies we use to tailor cloze tests that are efficient. We need to show the cloze tests "who is boss" by shaping them to our language testing purposes. In short, we need to tailor our cloze.

REFERENCES

Alderson, J. C. (1978). A study of the cloze procedure with native and non-native speakers of English (doctoral dissertation, University of Edinburgh).

Alderson, J. C. (1979a). Scoring procedures for use on cloze tests. In C. A. Yorio, K. Perkins, & J. Schachter (Eds.), *On TESOL '79* (pp. 193-205). Washington, DC: TESOL.

Alderson, J. C. (1979b). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly, 13*(2), 219-227.

Alderson, J. C. (1980). Native and non-native speaker performance on cloze tests. *Language Learning, 30*, 59-76.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*, 535-555.

Bormuth, J. R. (1965). Validities of grammatical and semantic classifications of cloze test scores. In J. A. Figurel (Ed.), *Reading and inquiry* (pp. 283-285). Newark, DE: International Reading Associates.

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension tests scores. *Journal of Reading, 10*, 291-299.

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *Modern Language Journal, 64*, 311-317.

Brown, J. D. (1983). A closer look at cloze: Validity and reliability. In J. W. Oller, Jr. (Ed.) *Issues in Language Testing Research* (pp. 237-250). Rowley, MA: Newbury House.

Brown, J. D. (1984). A cloze is a cloze is a cloze? In J. Handscombe, R. A. Orem, & B. P. Taylor (Eds.), *On TESOL '83* (pp. 109-119). Washington, DC: TESOL.

Brown, J. D. (1986). Cloze procedure: a tool for teaching reading. *TESOL Newsletter 20*(5), 1 & 7.

Brown, J. D. (1988a). *Understanding research in second language learning: A teacher's guide to statistics and research design*. Cambridge: Cambridge University.

Brown, J. D. (1988b). Tailored cloze: improved with classical item analysis techniques. *Language Testing, 5*, 19-31.

Brown, J. D. (1989). Cloze item difficulty. *JALT Journal, 11*, 46-67.

Brown, J. D. (1993). What are the characteristics of *natural* cloze tests? *Language Testing, 10*, 93-116.

Brown, J. D. (1994). A closer look at cloze: Validity and reliability. In J. W. Oller, Jr. & J. Jonz (Eds.), *Cloze and coherence*. Lewisburg, PA: Associated University Presses. [Reprinted by permission from the original: Brown, J. D. (1983).]

Brown, J. D. (1998). Statistics Corner: Questions and answers about language testing statistics (Reliability and cloze test length). *Shiken: JALT Testing & Evaluation SIG Newsletter*, *2*(2), 19-22. Also retrieved March 1, 2003 from the World Wide Web: http://www.jalt.org/test/bro_3.htm

Brown, J. D., Yamashiro, A. D., & Ogane, E. (1999). Tailored cloze: Three ways to improve cloze tests. *University of Hawaii Working Papers in ESL, 17*(2), 107-129.

Brown, J. D., Yamashiro, A. D., & Ogane, E. (2001). The Emperor's new cloze: Strategies for revising cloze tests. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development* (pp. 143-161). Honolulu, HI: University of Hawai'i Press.

Chavez-Oller, M. A., Chihara, T., Weaver, K. A., & Oller, J. W., Jr. (1985). When are cloze items sensitive to constraints across sentences? *Language Learning, 35*, 181-206.

Chihara, T., Oller, J. W., Jr., Weaver, K. A., & Chavez-Oller, M. A. (1977). Are cloze items sensitive to constraints across sentences? *Language Learning, 27*, 63-73.

Cohen, A. D. (1980). *Testing language ability in the classroom*. Rowley, MA: Newbury House.

Conrad, C. (1970). *The cloze procedure as a measure of English proficiency.* Unpublished master's thesis, University of California Los Angeles.

Crawford, A. (1970). *The cloze procedure as a measure of reading comprehension of elementary level Mexican-American and Anglo-American children.* Unpublished doctoral dissertation, University of California Los Angeles.

Cronbach, L. J. (1970). *Essentials of psychological testing* (pp. 145-146). New York: Harper and Row.

Darnell, D. K. (1970). Clozentropy: a procedure for testing English language proficiency of foreign students. *Speech Monographs, 37*, 36-46.

Fry, E. (1985). *The NEW reading teacher's book of lists*. Englewood Cliffs, NJ: Prentice-Hall.

Gaies, S. J. (1980). T-unit analysis in second language research: Applications, problems and limitations. *TESOL Quarterly, 14*, 53-60.

Gallant, R. (1965). Use of cloze tests as a measure of readability in the primary grades. In J. A. Figurel (Ed.), *Reading and inquiry* (pp. 286-287). Newark, Delaware: International Reading Associates.

Goodman, K. S. (1967). Reading: A Psychological guessing game. *Journal of the Reading Specialist, 6*, 126-135.

Hinofotis, F. B. (1980). Cloze as an alternative method of ESL placement and proficiency testing. In J. W. Oller Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 121-128). Rowley, MA: Newbury House.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. Champaign, IL: National Council of Teachers of English.

Irvine, P., P. Atai and J.W. Oller Jr. (1974). Cloze, dictation, and the Test of English as a Foreign Language. *Language Learning, 24*, 245-252.

Jonz, J. (1976). Improving on the basic egg: the M-C cloze. *Language Learning, 26*, 255-256.

Jonz, J. (1987). Textual cohesion and second language comprehension. *Language Learning, 37*, 409-38.

Klare, G. P. (1984). Readability. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 681-744). NY: Longman.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometirika, 2*, 151-160.

Lorge, I. (1959). *The Lorge formula for estimating difficulty of reading materials*. New York: Columbia Teachers College.

Lotus. (1985). *1-2-3*. Cambridge, MA: Lotus Development.

Mardia, K. V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika, 58*(1), 105-121.

Markham, P. L. (1985). The rational deletion cloze and global comprehension in German. *Language Learning, 35*, 423-430.

Mullen, K. (1979). More on cloze tests as tests of proficiency in English as a second language. In E.J. Briere and F.B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp. 21-32). Washington, DC: TESOL.

Oller, J. W. Jr. (1972a). Dictation as a test of ESL proficiency. In H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings* (pp. 346-354). New York: McGraw-Hill.

Oller, J. W. Jr. (1972b). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modern Language Journal, 56*, 151-158.

Oller, J. W. Jr. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

Oller, J. W. Jr., & Inal, N. (1971). A cloze test of English prepositions. *TESOL Quarterly*, *5*, 315-326.

Pike, L. W. (1973). *An evaluation of present and alternative item formats for use in the Test of English as a Foreign Language*. Princeton, NJ: Educational Testing Service.

Porter, D. (1983). The effect of quantity of context on the ability to make linguistic predictions: A flaw in a measure of general proficiency. In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 63-74). London: Academic Press.

Ruddell, R. B. (1964). A study of the cloze comprehension technique in relation to structurally controlled reading material. *Improvement of Reading Through Classroom Practice, 9*, 298-303.

Smith, F. (1978). *Comprehension and learning.* New York: Holt, Rinehart, & Winston.

Stubbs, J. B., & Tucker, G. R. (1974). The cloze test as a measure of ESL proficiency for Arab students. *Modern Language Journal, 58*, 239-241.

Tabachnick, B. G., & Fidell, L. S. (2001). Using multivariate statistics (4th ed.). Boston, MA: Allyn & Bacon.

Taylor, W. L. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly, 30*, 414-438.

Thorndike, E.L., & Lorge, I. (1959). *The teacher's word book of 30,000 words*. New York: Columbia Teachers' College.

APPENDIX A:
EXAMPLE CLOZE TEST

Name _____ ***Your nationality*** _____
(Last) ***(First)***

How much time have you spent in English speaking countries?

Directions: Fill in one word in each blank. You may write directly on the test.
Example: The girl was walking down the street when she stepped on some ice and fell (ex. 1) ***down***
.

The Science of Automatic Control (Form E)

The science of automatic control depends on certain common principles by which an organism, machine, or system regulates itself. Many historical developments up to the present day have helped to identify these principles.

For hundreds of years (1)_____ were many examples of automatic control systems, but no connections (2)_____ recognized among them. A very early example was a device (3)_____ windmills designed to keep their sails facing into the wind. (4)_____ consisted simply of a miniature windmill, which rotated the whole (5)_____ to face in any direction. The small mill was (6)_____ right angles to the main one, and whenever the latter faced (7)_____ the wrong direction, the wind caught the small mill's sails (8)_____ rotated the main mill to the correct position. Other automatic (9)_____ mechanisms were invented with the development of steam power: first (10)_____ engine governor, and then the steering engine controller…
(continues for a total of 30 items)

APPENDIX B:
EXAMPLE CLOZE ANSWER KEY

Form E - exact answers

1. THERE
2. WERE
3. ON
4. IT
5. MILL
6. AT
7. IN
8. AND
9. CONTROL
10. THE

… (continues for a total of 30 items)

APPENDIX C:

ITEM FACILITY AND DISCRIMINATION ON

FIVE EX SCORED CLOZE TESTS

FOR HIGH AND LOW PROFICIENCY STUDENTS

(discrimination = .30 or higher)*

Item	Word	P of S	C/F	LoGpIF	LoGpID	LoGp	HiGp	HiGpIF	HiGpID
A01	many	Adj	C	0.20	0.54	*	*	0.72	0.40
A02	among	prep	F	0.10	0.31	*		0.24	0.20
A03	designed	verb	C	0.08	0.23			0.07	0.10
A04	simply	Adv	C	0.03	0.08			0.07	0.10
A05	face	verb	C	0.03	0.08			0.45	0.00
A06	to	prep	F	0.03	0.08		*	0.48	0.50
A07	wrong	Adj	C	0.03	0.08		*	0.34	0.40
A08	the	Art	F	0.08	0.15		*	0.62	0.40
A09	were	verb	C	0.18	0.31	*	*	0.76	0.50
A10	governor	noun	C	0.03	0.08			0.00	0.00
A11	rudder	noun	C	0.00	0.00			0.00	0.00
A12	others	pron	C	0.03	0.08		*	0.24	0.40
A13	to	prep	F	0.28	0.69	*		0.62	0.00
A14	rapid	adj	C	0.00	0.00			0.07	0.00
A15	solutions	noun	C	0.00	0.00			0.07	0.20
A16	automatic	adj	C	0.15	0.23		*	0.55	0.50
A17	temperature	noun	C	0.00	0.00			0.03	0.10
A18	systems	noun	C	0.03	0.08		*	0.28	0.40
A19	signals	noun	C	0.03	0.08			0.00	0.00
A20	aided	verb	C	0.00	0.00			0.03	0.10
A21	to	inf	F	0.25	0.38	*	*	0.76	0.70
A22	recognized	verb	C	0.00	0.00		*	0.14	0.30
A23	depend	verb	C	0.05	0.15			0.10	0.20
A24	human	adj	C	0.00	0.00			0.21	0.20
A25	give	verb	C	0.03	0.08			0.14	0.10
A26	human	adj	C	0.00	0.00			0.10	0.10
A27	in	prep	F	0.00	0.00		*	0.41	0.70
A28	the	art	F	0.00	0.00			0.48	0.10
A29	and	conj	F	0.00	0.00		*	0.59	0.30
A30	regularly	adv	C	0.00	0.00			0.00	0.00
B01	examples	noun	C	0.05	0.15		*	0.10	0.30
B02	them	pron	C	0.13	0.38	*	*	0.73	0.50
B03	to	inf	F	0.53	0.46	*		0.93	0.20
B04	of	prep	F	0.18	0.31	*	*	0.37	0.60
B05	in	prep	F	0.11	0.23		*	0.23	0.30
B06	the	art	F	0.21	0.54	*	*	0.70	0.40

B07	direction	noun	C	0.03	0.08		*	0.67	0.60
B08	main	adj	C	0.05	0.15		*	0.40	0.30
B09	invented	verb	C	0.03	0.08			0.13	0.20
B10	and	conj	F	0.16	0.38	*	*	0.60	0.60
B11	in	prep	F	0.03	0.08		*	0.13	0.30
B12	constituted	verb	C	0.00	0.00			0.00	0.00
B13	about	adv	C	0.08	0.23		*	0.47	0.40
B14	technological	adj	C	0.00	0.00			0.00	0.00
B15	to	prep	F	0.00	0.00		*	0.37	0.50
B16	control	adj	C	0.37	0.38	*	*	0.63	0.60
B17	and	conj	F	0.16	0.15			0.70	0.10
B18	radios	noun	C	0.00	0.00			0.00	0.00
B19	historically	adv	C	0.00	0.00			0.00	0.00
B20	by	prep	F	0.00	0.00		*	0.20	0.40
B21	recall	verb	C	0.00	0.00			0.00	0.00
B22	yet	conj	F	0.00	0.00			0.00	0.00
B23	on	prep	F	0.37	0.69	*		0.93	0.10
B24	affairs	noun	C	0.00	0.00			0.00	0.00
B25	us	pron	C	0.03	0.08		*	0.50	0.30
B26	phenomena	noun	C	0.00	0.00			0.00	0.00
B27	understanding	verb	C	0.00	0.00		*	0.10	0.30
B28	human	adj	C	0.03	0.08		*	0.40	0.60
B29	booms	noun	C	0.00	0.00			0.03	0.10
B30	fluctuates	verb	C	0.00	0.00			0.00	0.00
C01	of	prep	F	0.18	0.46	*		0.93	0.10
C02	a	art	F	0.05	0.15		*	0.41	0.30
C03	keep	verb	C	0.03	0.08			0.03	0.10
C04	a	art	F	0.03	0.08		*	0.48	0.90
C05	any	adj	C	0.03	0.08			0.00	0.00
C06	main	adj	C	0.03	0.08			0.14	0.10
C07	the	art	F	0.15	0.46	*	*	0.72	0.60
C08	mill	noun	C	0.00	0.00			0.62	0.20
C09	with	prep	F	0.00	0.00			0.03	0.10
C10	then	adv	C	0.00	0.00		*	0.52	0.40
C11	correspondence	noun	C	0.00	0.00			0.00	0.00
C12	the	art	F	0.00	0.00		*	0.21	0.60
C13	fifty	adj	C	0.03	0.08		*	0.55	0.70
C14	development	noun	C	0.00	0.00			0.31	0.20
C15	these	adj	C	0.03	0.08		*	0.69	0.40
C16	devices	noun	C	0.00	0.00			0.03	0.00
C17	flow	noun	C	0.00	0.00			0.00	0.00
C18	required	verb	C	0.00	0.00			0.00	0.00
C19	then	adv	C	0.03	0.08			0.03	0.10
C20	related	verb	C	0.00	0.00			0.00	0.00
C21	that	conj	F	0.03	0.08		*	0.34	0.50

C22	we	pron	C	0.00	0.00		*	0.52	0.50
C23	common	adj	C	0.10	0.23			0.31	0.10
C24	indeed	adv	C	0.00	0.00			0.00	0.00
C25	new	adj	C	0.00	0.00			0.00	0.00
C26	the	art	F	0.00	0.00		*	0.59	0.50
C27	how	adv	C	0.00	0.00		*	0.41	0.70
C28	heart	noun	C	0.00	0.00		*	0.66	0.80
C29	and	conj	F	0.03	0.08		*	0.86	0.30
C30	between	prep	F	0.00	0.00		*	0.17	0.40
D01	were	verb	C	0.32	0.77	*		0.71	0.22
D02	recognized	verb	C	0.03	0.08			0.07	0.11
D03	windmills	noun	C	0.00	0.00		*	0.21	0.33
D04	consisted	verb	C	0.05	0.08			0.04	0.11
D05	to	inf	F	0.16	0.38	*	*	0.68	0.44
D06	angles	noun	C	0.00	0.00			0.07	0.22
D07	the	art	F	0.21	0.54	*	*	0.61	0.78
D08	rotated	verb	C	0.00	0.00			0.00	0.00
D09	mechanisms	noun	C	0.00	0.00			0.00	0.00
D10	engine	noun	C	0.03	-0.08			0.00	0.00
D11	ship's	noun	C	0.00	0.00			0.00	0.00
D12	few	adj	C	0.00	0.00		*	0.29	0.67
D13	up	prep	F	0.03	0.00		*	0.32	0.44
D14	however	conj	F	0.00	0.00			0.11	0.00
D15	the	art	F	0.08	0.15		*	0.39	0.67
D16	of	prep	F	0.24	0.46	*		0.36	0.22
D17	both	adj	C	0.00	0.00			0.04	0.11
D18	cooling	adj	C	0.11	0.15			0.25	0.00
D19	of	prep	F	0.00	0.00		*	0.54	0.44
D20	been	verb	C	0.13	0.38	*	*	0.75	0.44
D21	surprising	adj	C	0.00	0.00			0.00	0.00
D22	originally	adv	C	0.00	0.00			0.00	0.00
D23	systems	noun	C	0.03	0.08			0.14	0.11
D24	and	conj	F	0.05	0.15		*	0.86	0.33
D25	systems	noun	C	0.08	0.15			0.07	0.00
D26	and	conj	F	0.00	0.00		*	0.36	0.33
D27	helpful	adj	C	0.00	0.00		*	0.21	0.44
D28	how	adv	C	0.00	0.00		*	0.71	0.44
D29	slumps	noun	C	0.00	0.00			0.00	0.00
D30	Canada	noun	C	0.00	0.00			0.00	0.00
E01	there	pron	C	0.18	0.38	*		0.93	0.11
E02	were	verb	C	0.18	0.54	*	*	0.52	0.33
E03	on	prep	F	0.03	0.08			0.04	0.11
E04	it	pron	C	0.05	0.15			0.67	0.22
E05	mill	noun	C	0.03	0.08			0.00	0.00

E06	at	prep	F	0.00	0.00			0.00	0.00
E07	in	prep	F	0.18	0.15			0.37	0.22
E08	and	conj	F	0.03	0.08		*	0.70	0.44
E09	control	noun	C	0.53	0.54	*	*	0.78	0.33
E10	the	art	F	0.08	0.15		*	0.56	0.44
E11	a	art	F	0.03	0.08			0.04	0.00
E12	a	art	F	0.13	0.23			0.33	0.11
E13	control	noun	C	0.39	0.46	*	*	0.63	0.33
E14	decades	noun	C	0.03	0.08			0.11	0.00
E15	problems	noun	C	0.03	0.08		*	0.37	0.67
E16	families	noun	C	0.03	0.08			0.00	0.00
E17	for	prep	F	0.03	0.08			0.41	0.11
E18	and	conj	F	0.21	0.38	*		0.93	0.11
E19	accuracy	noun	C	0.03	0.08			0.04	0.11
E20	has	verb	C	0.11	0.23			0.52	0.11
E21	seems	verb	C	0.03	0.08			0.07	0.22
E22	not	adv	C	0.00	0.00		*	0.48	0.56
E23	regulating	adj	C	0.00	0.00			0.00	0.00
E24	nature	noun	C	0.00	0.00			0.19	0.22
E25	control	adj	C	0.53	0.38	*	*	0.85	0.33
E26	natural	adj	C	0.00	0.00			0.11	0.22
E27	very	adv	C	0.00	0.00			0.48	0.11
E28	upright	adv	C	0.00	0.00		*	0.11	0.33
E29	from	prep	F	0.21	0.31	*	*	0.37	0.44
E30	of	prep	F	0.45	0.54	*	*	0.85	0.33

APPENDIX D:

ITEM FACILITY AND DISCRIMINATION ON

FIVE AC SCORED CLOZE TESTS

FOR HIGH AND LOW PROFICIENCY STUDENTS

(* discrimination = .30 or higher)

Item	Original Word	P of S	C/F	LoGpIF	LoGpID	LoGp	HiGp	HiGpIF	HiGpID
A01	many	adj	C	0.20	0.54	*		0.86	0.10
A02	among	prep	F	0.10	0.31	*		0.24	0.10
A03	designed	verb	C	0.08	0.23		*	0.31	0.50
A04	simply	adv	C	0.03	0.08		*	0.24	0.50
A05	face	verb	C	0.03	0.08			0.45	-0.10
A06	to	prep	F	0.03	0.08		*	0.48	0.50
A07	wrong	adj	C	0.03	0.08		*	0.34	0.50
A08	the	art	F	0.08	0.00		*	0.62	0.30
A09	were	verb	C	0.18	0.38	*	*	0.76	0.60
A10	governor	noun	C	0.03	0.08			0.00	0.00
A11	rudder	noun	C	0.00	0.00			0.00	0.00
A12	others	pron	C	0.03	0.08		*	0.24	0.60
A13	to	prep	F	0.28	0.62	*		0.83	0.20
A14	rapid	adj	C	0.03	0.08			0.17	0.20
A15	solutions	noun	C	0.00	0.00		*	0.28	0.40
A16	automatic	adj	C	0.15	0.15		*	0.55	0.50
A17	temperature	noun	C	0.03	0.08			0.03	0.00
A18	systems	noun	C	0.03	0.08		*	0.34	0.50
A19	signals	noun	C	0.03	0.08			0.03	0.10
A20	aided	verb	C	0.00	0.00			0.03	0.00
A21	to	inf	F	0.25	0.46	*	*	0.76	0.60
A22	recognized	verb	C	0.00	0.00		*	0.24	0.60
A23	depend	verb	C	0.08	0.15		*	0.17	0.30
A24	human	adj	C	0.00	0.00			0.21	0.10
A25	give	verb	C	0.03	0.08			0.24	0.00
A26	human	adj	C	0.00	0.00			0.10	0.00
A27	in	prep	F	0.10	0.15		*	0.86	0.30
A28	the	art	F	0.03	0.08		*	0.76	0.30
A29	and	conj	F	0.03	0.00		*	0.59	0.50
A30	regularly	adv	C	0.00	0.00		*	0.14	0.30
B01	examples	noun	C	0.05	0.15		*	0.60	0.50
B02	them	pron	C	0.13	0.38	*	*	0.80	0.40
B03	to	inf	F	0.53	0.38	*		0.93	0.10
B04	of	prep	F	0.21	0.31	*	*	0.37	0.40
B05	in	prep	F	0.16	0.31	*		0.43	0.20
B06	the	art	F	0.21	0.46	*	*	0.70	0.30

B07	direction	noun	C	0.03	0.08		*	0.77	0.30
B08	main	adj	C	0.13	0.08		*	0.53	0.50
B09	invented	verb	C	0.03	0.08			0.13	0.10
B10	and	conj	F	0.16	0.38	*	*	0.60	0.70
B11	in	prep	F	0.03	0.08		*	0.13	0.30
B12	constituted	verb	C	0.00	0.00			0.00	0.00
B13	about	adv	C	0.08	0.15		*	0.60	0.60
B14	technological	adj	C	0.00	0.00			0.10	0.10
B15	to	prep	F	0.13	0.38	*	*	0.73	0.60
B16	control	adj	C	0.37	0.38	*	*	0.63	0.60
B17	and	conj	F	0.16	0.23			0.70	0.20
B18	radios	noun	C	0.00	0.00			0.00	0.00
B19	historically	adv	C	0.00	0.00			0.00	0.00
B20	by	prep	F	0.00	0.00		*	0.20	0.30
B21	recall	verb	C	0.00	0.00		*	0.60	0.50
B22	yet	conj	F	0.05	0.15			0.20	-0.10
B23	on	prep	F	0.37	0.77	*		0.97	0.10
B24	affairs	noun	C	0.05	0.00			0.17	0.00
B25	us	pron	C	0.03	0.08			0.53	0.10
B26	phenomena	noun	C	0.00	0.00			0.03	0.10
B27	understanding	verb	C	0.00	0.00		*	0.17	0.30
B28	human	adj	C	0.03	0.08		*	0.57	0.60
B29	booms	noun	C	0.00	0.00			0.03	0.10
B30	fluctuates	verb	C	0.00	0.00			0.00	0.00
C01	of	prep	F	0.28	0.46	*		0.93	0.10
C02	a	art	F	0.05	0.08		*	0.66	0.60
C03	keep	verb	C	0.03	0.08			0.03	0.10
C04	a	art	F	0.03	0.08		*	0.55	0.90
C05	any	adj	C	0.03	0.08			0.14	0.20
C06	main	adj	C	0.03	0.08		*	0.66	0.40
C07	the	art	F	0.18	0.46	*	*	0.72	0.60
C08	mill	noun	C	0.08	0.15		*	0.76	0.40
C09	with	prep	F	0.00	0.00			0.14	0.10
C10	then	adv	C	0.15	0.31	*	*	0.86	0.40
C11	correspondence	noun	C	0.00	0.00			0.00	0.00
C12	the	art	F	0.00	0.00		*	0.21	0.50
C13	fifty	adj	C	0.03	0.08		*	0.55	0.60
C14	development	noun	C	0.05	0.08		*	0.59	0.50
C15	these	adj	C	0.13	0.38	*	*	0.72	0.50
C16	devices	noun	C	0.13	0.08		*	0.55	0.50
C17	flow	noun	C	0.03	0.08			0.28	0.00
C18	required	verb	C	0.03	0.08		*	0.66	0.60
C19	then	adv	C	0.03	0.08			0.03	0.10
C20	related	verb	C	0.10	0.31	*	*	0.59	0.50

C21	that	conj	F	0.03	0.08		*	0.38	0.50
C22	we	pron	C	0.00	0.00		*	0.69	0.60
C23	common	adj	C	0.21	0.54	*		0.72	-0.10
C24	indeed	adv	C	0.03	0.00			0.17	0.20
C25	new	adj	C	0.00	0.00			0.07	0.10
C26	the	art	F	0.00	0.00		*	0.66	0.50
C27	how	adv	C	0.00	0.00		*	0.76	0.50
C28	heart	noun	C	0.00	0.00		*	0.66	0.70
C29	and	conj	F	0.03	0.08		*	0.86	0.40
C30	between	prep	F	0.00	0.00		*	0.17	0.50
D01	were	verb	C	0.32	0.69	*	*	0.79	0.44
D02	recognized	verb	C	0.05	0.08		*	0.61	0.78
D03	windmills	noun	C	0.00	0.00			0.32	0.11
D04	consisted	verb	C	0.05	0.00			0.04	0.00
D05	to	inf	F	0.16	0.38	*	*	0.68	0.67
D06	angles	noun	C	0.00	0.00			0.07	0.22
D07	the	art	F	0.21	0.54	*	*	0.64	0.56
D08	rotated	verb	C	0.00	0.00		*	0.46	0.44
D09	mechanisms	noun	C	0.47	0.23		*	0.64	0.44
D10	engine	noun	C	0.03	0.00			0.00	0.00
D11	ship's	noun	C	0.00	0.00			0.00	0.00
D12	few	adj	C	0.00	0.00		*	0.29	0.67
D13	up	prep	F	0.03	0.00		*	0.32	0.56
D14	however	conj	F	0.00	0.00			0.18	0.22
D15	the	art	F	0.08	0.15		*	0.57	0.33
D16	of	prep	F	0.24	0.31	*		0.43	0.11
D17	both	adj	C	0.03	0.08		*	0.25	0.56
D18	cooling	adj	C	0.13	0.15			0.39	0.00
D19	of	prep	F	0.00	0.00		*	0.61	0.56
D20	been	verb	C	0.13	0.38	*	*	0.75	0.44
D21	surprising	adj	C	0.00	0.00			0.04	0.11
D22	originally	adv	C	0.00	0.00			0.07	0.00
D23	systems	noun	C	0.03	0.08			0.14	0.11
D24	and	conj	F	0.05	0.15		*	0.86	0.33
D25	systems	noun	C	0.11	0.15			0.18	0.11
D26	and	conj	F	0.03	0.08			0.36	0.11
D27	helpful	adj	C	0.00	0.00		*	0.61	0.56
D28	how	adv	C	0.00	0.00			0.93	0.22
D29	slumps	noun	C	0.00	0.00			0.04	0.11
D30	Canada	noun	C	0.00	0.00			0.07	0.22
E01	there	pron	C	0.18	0.31	*		0.93	0.22
E02	were	verb	C	0.18	0.38	*	*	0.52	0.33
E03	on	prep	F	0.05	0.08			0.04	0.11
E04	it	pron	C	0.05	0.15		*	0.70	0.33

E05	mill	noun	C	0.03	0.08			0.30	0.00
E06	at	prep	F	0.00	0.00			0.04	0.11
E07	in	prep	F	0.34	0.46	*		0.74	0.00
E08	and	conj	F	0.05	0.08			0.89	0.22
E09	control	noun	C	0.53	0.54	*	*	0.78	0.33
E10	the	art	F	0.08	0.08		*	0.59	0.89
E11	a	art	F	0.03	0.08		*	0.56	0.33
E12	a	art	F	0.13	0.23			0.33	0.22
E13	control	noun	C	0.39	0.38	*		0.63	0.22
E14	decades	noun	C	0.03	0.08			0.11	0.11
E15	problems	noun	C	0.03	0.08		*	0.37	0.78
E16	families	noun	C	0.03	0.08			0.11	0.22
E17	for	prep	F	0.03	0.08			0.56	0.22
E18	and	conj	F	0.24	0.38	*		0.93	0.11
E19	accuracy	noun	C	0.03	0.08		*	0.11	0.33
E20	has	verb	C	0.11	0.23			0.52	0.22
E21	seems	verb	C	0.16	0.31	*		0.78	0.22
E22	not	adv	C	0.00	0.00		*	0.48	0.33
E23	regulating	adj	C	0.13	0.23		*	0.26	0.33
E24	nature	noun	C	0.03	0.08			0.59	0.00
E25	control	adj	C	0.53	0.31	*		0.85	0.11
E26	natural	adj	C	0.00	0.00		*	0.19	0.44
E27	very	adv	C	0.00	0.00		*	0.67	0.33
E28	upright	adv	C	0.00	0.00		*	0.30	0.56
E29	from	prep	F	0.24	0.38	*	*	0.56	0.33
E30	of	prep	F	0.45	0.62	*	*	0.85	0.33

James Dean Brown
Department of Second Language Studies
University of Hawai‘i at Mānoa
1890 East-West Road
Honolulu, HI 96822

brownj@hawaii.edu